SIAM REVIEW
Vol. 62, No. 2, pp. 483–508

Constructing Least-Squares Polynomial Approximations[*]

Ling Guo[†]
Akil Narayan[‡]
Tao Zhou[§]

Abstract. Polynomial approximations constructed using a least-squares approach form a ubiquitous technique in numerical computation. One of the simplest ways to generate data for least-squares problems is with random sampling of a function. We discuss theory and algorithms for stability of the least-squares problem using random samples. The main lesson from our discussion is that the intuitively straightforward ("standard") density for sampling frequently yields suboptimal approximations, whereas sampling from a non-standard density, called the induced distribution, yields near-optimal approximations. We present a recent theory that demonstrates why sampling from the induced distribution is optimal and provide several numerical experiments that support the theory. Software is also provided that reproduces the figures in this paper.

Key words. least-squares approximations, optimal sampling, polynomial approximations

AMS subject classifications. 41A10, 41A25, 41A65, 62E17, 93E24

DOI. 10.1137/18M1234151

1. Introduction.

Many applications require the construction of approximations to a given function f. When f is complicated or expensive to evaluate, one typically transforms evaluations of f on a grid into an approximation g. A simple example of this procedure is univariate polynomial interpolation, where f is sampled at M distinct points and g is subsequently built as a degree-$(M-1)$ polynomial that uniquely interpolates the data from f. Other examples of approximation procedures are common as well; here we will focus on one of the simplest approximation techniques: polynomial approximation via discrete least squares. In particular, we will focus on the case when the abscissae on which f is evaluated are randomly drawn, which has advantages in both theory and practice.

[*] Received by the editors December 19, 2018; accepted for publication (in revised form) July 17, 2019; published electronically May 7, 2020.
https://doi.org/10.1137/18M1234151

Funding: The work of the first author was supported by the NSF of China through grant 11671265. The work of the second author was partially supported by AFOSR FA9550-15-1-0467 and NSF DMS-1720416. The work of the third author was partially supported by the NSF of China (under grants 11822111, 11688101, 91630312, 91630203, 11571351, and 11731006), the Science Challenge Project (TZ2018001), the National Key Basic Research Program (2018YFB0704304), NCMIS, and the Youth Innovation Promotion Association (CAS).

[†] Department of Mathematics, Shanghai Normal University, Shanghai, China 200234 (lguo@shnu.edu.cn).

[‡] Department of Mathematics, and Scientific Computing and Imaging (SCI) Institute, University of Utah, Salt Lake City, UT 84112 (akil@sci.utah.edu).

[§] LSEC, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, Beijing, China 100190 (tzhou@lsec.cc.ac.cn).

While the construction of approximations in one dimension is well studied and understood (see, e.g., [55]), the task of building such an approximation when f is a multivariate function is substantially harder. Such problems arise, for example, in parametric uncertainty quantification problems, where analyzing dependence of a mathematical model on random input parameters is transformed into a problem of forming a polynomial approximation in a high-dimensional space [57, 47, 40, 32, 19, 53, 60, 31].

Many methodologies for performing function approximation in high-dimensional spaces utilize quadrature rules or interpolation grids, but construction of such grids is a difficult problem in high dimensions, and many standard constructions yield a grid size so large that one cannot afford many evaluations of f. It is in such situations when discrete least-squares procedures, particularly those where samples are unstructured and *randomly* chosen, can be beneficial. This is the main point that we wish to drive home in this paper. In particular, our focus is on the demonstration of the theory and practice of least-squares construction of polynomial approximations using random samples in multiple dimensions. There is a plethora of literature that studies least-squares statistical regression estimation procedures (especially in the noisy data case), which we briefly discuss.

1.1. Comparison to Statistical Regression. The statistical community has long developed regression tools for generating parametric models that furnish relationships between variables. These tools are not the focus of this paper, but it is appropriate for us to briefly describe the similarities and differences between the approach of this paper and that of regression analysis. We refer the reader to [20, 36, 28] for more proper, complete descriptions of regression techniques. We also note that topics such as learning theory [3, 4] and optimal experimental design [21, 34, 9] consider similar problems.

Many regression techniques aim to fit observational data to prescribed models, and determining model parameters via ordinary least squares is one standard computational approach. When the noise in the observed data is independent but not stationary (heteroscedasticity), then weights can be introduced in the least-squares objective to "whiten" the impact of each sample on the objective function. Once a model is built, regression is often concerned with topics such as understanding biasing properties of the model parameters, determining variable influence, and understanding asymptotic convergence properties of the estimator.

The *computational* approach we take in this paper is weighted least squares and hence is quite similar to a weighted least-squares regression procedure. However, in contrast to the viewpoint above, we are interested in building an *approximation* to a true underlying function. In this sense, in general we expect convergence of the model to some "best" model and not to the actual underlying function as the number of samples is increased. The data we generate is not noisy, and hence a statistical interpretation of the data may not be appropriate. Finally, one can mathematically define a "best" model, but a model built from finite data may not match the best one. We are primarily interested in understanding when and how we can obtain a model that is close to the best one.[1]

1.2. Scope of This Paper. In statistical parlance our main computational objective is as follows: Given observational data, we seek to build a model whose parameters are chosen via least-squares regression. The goal of this paper is similar in philosophy to statistical regression, but we reword the previous statement in mathematical

[1] We make this more precise and explicit in section 5.1.

language, whose connotations more accurately reflect the context: Given function samples, we seek to build a mathematical model from a prescribed vector space, and the expansion coefficients of this model are determined by a weighted least-squares method.

In this sense, we assume that the underlying function f is a member of a larger (infinite-dimensional) vector space, but that the model is finite-dimensional. (For example, a general analytic function can be approximated by a finite-degree polynomial model.) In this situation the function samples are indeed observational data, but are not noisy so that this is not the setup of standard statistical regression theory. Thus, large-data asymptotics of statistical regression are not the appropriate viewpoint. Instead, one expects in the large-sample limit that the approximation g converges to a function that is the *projection* of f onto the finite-dimensional model space. This projection is the "best" model mentioned at the end of section 1.1. A natural desire then is to understand how much data is required so that the predicted model is close to this projection.

The perspective above is the main focus of this paper. The approaches we discuss are more common in the computational mathematics discipline, in particular in the numerical analysis community. We consider the case when the function f from which data is gathered is deterministic, but the particular abscissa at which f is evaluated to generate data is random. We compute coefficients of an approximation via least squares on the function samples. Thus, randomness enters via the sampling process but does not correspond to noise or error in the data values.

When constructing least-squares approximations with randomly generated samples, the large-sample limit should yield an approximation that matches the best model to a function. Of course, in practice one wishes to reach this limiting case with as few samples as possible. Recent work in the literature has shown that sampling from "standard" densities requires a relatively large number of samples, whereas using samples from so-called *induced* distributions yields stable least-squares problems with a near-optimal (minimal) number of samples. The main goal of this paper is to illustrate the theory of such results and to demonstrate through numerical experiments that sampling from induced distributions can substantially improve the quality of approximation in least-squares problems.

The theory which gives rise to this paper was first developed in [14, 16, 45], and algorithms for sampling from nonstandard multivariate distributions that we employ are available in [42, 41]. In particular, MATLAB code that reproduces Figures 1–6 in this paper is available from [43]. A brief outline of this paper is as follows:

- Section 2: Formal statement of the problem
- Section 3: Approximations in one dimension with (deterministic) quadrature rules
- Section 4: Notation for the multivariate problem
- Section 5: Discussion of weighted least squares with finite data, and asymptotic convergence properties
- Section 6: Preasymptotic stability analysis for the least-squares procedure
- Section 7: Presentation of optimal (induced) sampling distributions
- Section 8: Discussion of induced distribution behavior for large polynomial degree
- Section 9: Demonstration of procedures on prediction for a parametric partial differential equation

- Sections 10 and 11: Conclusion and discussion of existing generalizations and new directions for research

2. Problem Statement. For $D \subset \mathbb{R}^d$ a domain, let $w : D \to [0, \infty)$ be a weight function, and consider the following space of w-weighted square-integrable real-valued functions over D:

$$L^2_w = L^2_w(D) = \left\{ u : D \to \mathbb{R} \;\middle|\; \int_D u^2(x) w(x) \mathrm{d}x < \infty \right\}.$$

The space L^2_w is a Hilbert space with an inner product and norm defined, respectively, as

$$\langle u, v \rangle \coloneqq \int_D u(x)v(x)w(x)\mathrm{d}x, \qquad \|u\|^2 \coloneqq \langle u, v \rangle .$$

To simplify some notation later, we will assume that w is a probability density function, i.e., that $\|1\| = 1$. This is not a particularly strong assumption since it is essentially equivalent to requiring that constant functions are in L^2_w. The following examples illustrate common choices for w and D.

EXAMPLE 2.1 (approximations on hypercubes). *Let the dimension $d \geq 1$ be fixed, and consider the hypercube $D = [-1, 1]^d$ with weight function $w(x) = 2^{-d}$ for $x \in D$, and $w(x) = 0$ otherwise.*

EXAMPLE 2.2 (approximations on $\mathbb{R}^d$). *Let the dimension $d \geq 1$ be fixed, and consider the region $D = \mathbb{R}^d$ with weight function $w(x) = (2\pi)^{-d} \exp(-\|x\|_2^2)$ for $x \in D$, where $\|x\|_2^2 = \sum_{j=1}^d x_j^2$.*

Given some $f \in L^2_w$, we are interested in building approximations to f. Specifically, we are interested in building *polynomial* approximations to f. For computational purposes, we can only construct approximations from a finite-dimensional subspace of L^2_w. Let V be this subspace, having dimension N. Assume that $v_1, v_2, \ldots, v_N$ is an L^2_w-orthonormal basis for V, i.e., that

$$\langle v_j, v_k \rangle = \delta_{j,k}, \qquad 1 \leq j, k \leq N, \tag{2.1}$$

where $\delta_{j,k}$ is the Kronecker delta function. The best possible approximation to a given $f \in L^2_w$ is the orthogonal projection onto V given by

$$f_N(x) \coloneqq \sum_{n=1}^N \widehat{f}_n v_n(x), \qquad \widehat{f}_n \coloneqq \langle f, v_n \rangle . \tag{2.2}$$

As stated, f_N is the function from V that is closest to f as measured in the L^2_w norm:

$$f_N = \operatorname*{argmin}_{v \in V} \|f - v\| . \tag{2.3}$$

Our main goal in this paper is to discuss a computational strategy for computing an approximation to f_N using least-squares approximation with random samples.

3. Approximation with Polynomials and a One-Dimensional Example. We specialize the content of the previous section in two ways. First, we now consider approximation with polynomials, and second, we will consider an explicit example with univariate approximation. Our main discussion will revolve around the approximation of f_N defined in (2.2). The main message of this section is that for one-dimensional cases there are constructive, deterministic ways to approximate f_N, but that these procedures suffer from computational bottlenecks for multivariate approximation, which motivates a later study of least squares.

Our univariate assumption implies that $D \subset \mathbb{R}$. Consider the case where V is a polynomial subspace of finite dimension; in particular, we will take the space of polynomials of degree $N-1$ and lower for some fixed $N \in \mathbb{N}$:

$$V \coloneqq \operatorname{span}\left\{1, x, \ldots, x^{N-1}\right\}. \tag{3.1}$$

In order to fit our discussion into that of section 2, we need to assume (a) that V is a subspace of L^2_w, and (b) that there exists an L^2_w-orthonormal basis $\{v_n\}_{n=1}^N$ for V. The first assumption can be guaranteed if all nontrivial polynomials from V have finite, nonvanishing L^2_w norm, and under this condition the second assumption is satisfied by prescribing the functions v_n as members of classical orthogonal polynomial families.

Example 3.1. *Let $w(x)$ be a weight function on $D \subset \mathbb{R}$, and let the subspace V be as in* (3.1).

- *Let $w(x) = 1/2$ for $x \in D = [-1, 1]$ and let it vanish outside this interval. Then any nontrivial polynomial p satisfies $0 < \|p\| < \infty$. For any N defining V, we can take v_n as the degree-n (normalized) Legendre polynomial. Such a basis of V is L^2_w-orthonormal.*
- *Let $w(x) = \exp(-x^2)/\sqrt{\pi}$ for $x \in D = \mathbb{R}$. Then, again, any nontrivial polynomial p satisfies $0 < \|p\| < \infty$. For any N defining V, we can take v_n as the degree-n (normalized) Hermite polynomial. Such a basis of V is L^2_w-orthonormal.*
- *Let $w(x) = 1/(\pi(1+x^2))$ for $x \in D = \mathbb{R}$. Then $V \not\subset L^2_w$ for any $N \geq 2$ since the L^2_w norm $\|x^q\|$ is not finite for any $q \geq 1$.*

The last example illustrates that while our setup of polynomial approximation is general, it does not cover all cases that one might consider.

Ideally we would like to compute f_N defined in (2.2); unfortunately the inner products (integrals) defining the coefficients $\widehat{f}_n$ are usually not exactly computable so that we must resort to approximations. A simple but effective approach to computing such integrals is by use of an M-point quadrature rule,

$$\widehat{f}_n = \int_D f(x) v_n(x) w(x) \mathrm{d}x \approx \sum_{m=1}^{M} f(x_m) v_n(x_m) \lambda_m \eqqcolon c_n. \tag{3.2}$$

Here, $(x_m, \lambda_m)_{m=1}^M$ are quadrature nodes and weights, respectively. Once the coefficients c_n are computed, we form a new approximation g_N defined as

$$g_N(x) \coloneqq \sum_{n=1}^{N} c_n v_n(x) \in V. \tag{3.3}$$

Computing expansion coefficients via quadrature, as in (3.2), is an effective technique in one dimension where optimal quadrature rules are well studied. For example, a classical univariate quadrature rule is the w-Gaussian quadrature rule, which is an M-point rule $(x_m, \lambda_m)_{m=1}^M$ that integrates all polynomials up to degree $2M-1$ exactly:

$$\int_D p(x) w(x) \mathrm{d}x = \sum_{m=1}^{M} p(x_m) \lambda_m \qquad \forall \text{ polynomials } p \text{ such that } \deg p \leq 2M-1.$$

The generation of this rule (i.e., of the abscissae x_m and weights λ_m) can be accomplished via relatively simple numerical linear algebraic procedures involving recurrence coefficients for orthogonal polynomials. A modern reference detailing the theory and algorithms for Gaussian quadrature is [22, Chapter 1], though many other earlier texts discuss Gaussian quadrature as well.

Gaussian quadrature rules are a cornerstone of numerical analysis. A Gaussian rule is the *unique* quadrature rule of *optimal* accuracy in the sense that no other M-point quadrature rule achieves exactness over the same space of polynomials. Since this rule has optimal accuracy, it is an excellent candidate for performing the approximation (3.2). Before illustrating this with an example, we first define a metric that we will use to evaluate the accuracy of the approximation g_N. Since f_N is the L^2_w-orthogonal projection onto V, then $f - f_N$ is orthogonal to any function in V; in particular,

$$\langle f - f_N, g_N - f_N \rangle = 0, \qquad\qquad g_N - f_N \in V.$$

Thus, by the Pythagorean theorem,

$$\|f - g_N\|^2 = \|f - f_N\|^2 + \|f_N - g_N\|^2 .$$

Dividing both sides by $\|f - f_N\|^2$, we see that the error that g_N commits in approximating f relative to its best approximation f_N can be characterized by the nonnegative number

$$\eta_N := \frac{\|g_N - f_N\|}{\|f - f_N\|}, \tag{3.4}$$

so that

$$\|f - g_N\|^2 = (1 + \eta_N^2)\|f - f_N\|^2.$$

We assume throughout this paper that $\|f - f_N\| > 0$; if this were not true, then f would be exactly a function from V, which is rarely the case in practice. Values of η_N that are approximately 1 imply that the construction of g_N is well behaved: the error committed by the act of constructing an approximation g_N from V to f is comparable to the error committed by the best approximation f_N. Values of η_N that are much larger than 1 indicate that the construction of g_N is ill behaved. With all the pieces in place, we can now investigate the accuracy of g_N constructed using Gaussian quadrature.

EXAMPLE 3.2. *Consider the $d = 1$ case of Example 2.1, i.e., approximation on the interval $D = [-1, 1]$ with the uniform weight function $w(x) = 1/2$ for $x \in D$. With V the degree-$(N-1)$ polynomial space in (3.1), we take orthonormal basis elements v_n to be degree-$(n-1)$ normalized Legendre polynomials. We consider approximating a scaled and shifted "bump" function $f(x)$ defined by*

$$f(x) := B\left(1.5x - 0.2\right), \qquad B(x) := \begin{cases} \exp\left(-\frac{1}{1-x^2}\right), & |x| < 1, \\ 0, & |x| \geq 1. \end{cases} \tag{3.5}$$

The bump function B is a standard example of a function that is infinitely differentiable, but not analytic. For each N, we can compute approximations c_n to the coefficients $\widehat{f}_n$ using an N-point Gaussian quadrature rule as in (3.2). (Hence, we

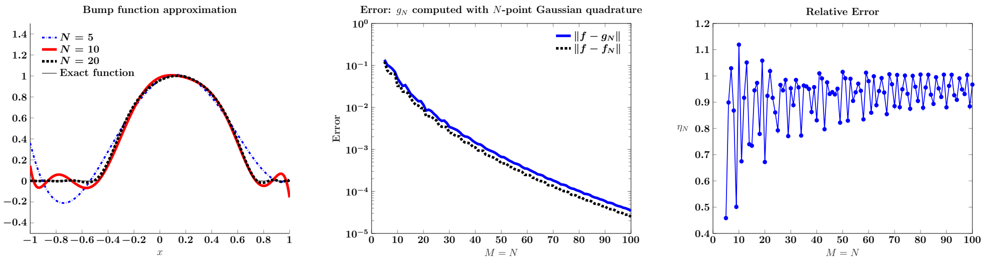

Fig. 1 *Approximation of the one-dimensional bump function f in (3.5) with polynomials using Gauss quadrature. Left: Original function and approximations g_N defined in (3.3) and (3.2) using an N-point Gauss quadrature rule to compute $\widetilde{f}_N$. Center: L^2_w errors for g_N and the best approximation f_N defined in (2.2). Right: Relative error metric η_N defined in (3.4). Values $\eta_N \lesssim 1$ indicate that the Gaussian quadrature strategy commits an error that is comparable to the best truncation error and thus that g_N is a near-best approximation.*

take $M = N$ here.) The left-hand pane of Figure 1 illustrates the function f along with the approximants g_N for a few values of N. The error of g_N as a function of N is shown in the center pane of Figure 1. Also shown is the best approximation error $\|f - f_N\|$. We see the hallmark of approximation of a smooth function by polynomials: high-order convergence with respect to the degree of approximation N.

Finally, the right-hand pane in Figure 1 shows the relative error metric η_N defined in (3.4). That this metric is $\mathcal{O}(1)$ for all N shown indicates that $\|f_N - g_N\|$, which is the discrepancy between the computed approximation g_N and the best approximation f_N, is of the same order of magnitude as the best approximation error $\|f - f_N\|$, and therefore g_N is a near-best approximation.

Note in the previous example that in order to construct g_N, we needed to specify N degrees of freedom (the coefficients c_n). We computed these using $M = N$ evaluations of f via a Gaussian quadrature rule. The situation is far more complicated in multiple dimensions.

4. Multivariate Notation.

We now wish to address the more difficult problem of computing multivariate polynomial approximations. To accomplish this we need to introduce multi-index notation. Let $\lambda \in \mathbb{N}_0^d = \{0, 1, \dots\}^d$ denote a multi-index. For $x \in \mathbb{R}^d$ and $\lambda \in \mathbb{N}_0^d$, we define x^λ in the standard way:

$$x = \left(x^{(1)}, x^{(2)}, \dots, x^{(d)}\right), \qquad \lambda = \left(\lambda^{(1)}, \lambda^{(2)}, \dots, \lambda^{(d)}\right), \qquad x^\lambda \coloneqq \prod_{j=1}^{d} \left[x^{(j)}\right]^{\lambda^{(j)}}.$$

We also define ℓ^p norms of multi-indices in the standard way:

$$\|\lambda\|_\infty \coloneqq \max_j \lambda^{(j)}, \qquad \|\lambda\|_p^p \coloneqq \sum_{j=1}^{d} \left[\lambda^{(j)}\right]^p, \qquad 0 < p < \infty.$$

An appropriate way to define polynomial spaces in several dimensions is to first identify a set of multi-indices and to subsequently define a polynomial space as the span of the corresponding monomials,

$$\Lambda = \{\lambda_1, \dots, \lambda_N\} \subset \mathbb{N}_0^d, \qquad V(\Lambda) \coloneqq \operatorname{span}\left\{x^\lambda \;\middle|\; \lambda \in \Lambda\right\}. \tag{4.1}$$

The subspace $V(\Lambda)$ has dimension N. In one dimension, if N is fixed and given, the choice of Λ is fairly straightforward (equaling $\{0, 1, \ldots, N-1\}$), but in multiple dimensions there are multiple "reasonable" choices for Λ. For example, given $k \geq 0$, there are ℓ^p-balls in $\mathbb{N}_0^d$,

$$\Lambda_p(k) \coloneqq \left\{ \lambda \in \mathbb{N}_0^d \;\middle|\; \|\lambda\|_p \leq k \right\}, \qquad\qquad 0 < p \leq \infty,$$

or the hyperbolic cross (HC) spaces

$$\Lambda_{\mathrm{HC}}(k) \coloneqq \left\{ \lambda \in \mathbb{N}_0^d \;\middle|\; \|\log(\lambda + 1)\|_1 \leq \log(k+1) \right\}, \tag{4.2a}$$

where $\log(\lambda)$ and $\lambda + 1$ are elementwise operations. Some important specializations of the ℓ^p-balls are the total degree (TD), tensor product (TP), and Euclidean degree (ED) spaces,

$$\Lambda_{\mathrm{TD}}(k) \coloneqq \Lambda_1(k), \qquad \Lambda_{\mathrm{TP}}(k) \coloneqq \Lambda_\infty(k), \qquad \Lambda_{\mathrm{ED}}(k) \coloneqq \Lambda_2(k). \tag{4.2b}$$

In one dimension, all of these spaces are the same: the space of degree-k polynomials. In the multivariate setting, they are different and each has its own uses, advantages, and shortcomings. Fix $d \geq 1$; for large "degree" $k \geq 1$, we have the following ordering for the dimension $N = |\Lambda|$ of these spaces, along with their asymptotics:

$$\begin{array}{ccccccc}
|\Lambda_{\mathrm{HC}}(k)| & < & |\Lambda_{\mathrm{TD}}(k)| & < & |\Lambda_{\mathrm{ED}}(k)| & < & |\Lambda_{\mathrm{TP}}(k)|, \\
\wr & & \wr & & \wr & & \parallel \\
(k+1)\log(k+1)^{d-1} & < & \frac{(k+1)^d}{\Gamma(d+1)} & < & \frac{1}{\Gamma\left(\frac{d}{2}+1\right)} \left[\frac{\sqrt{\pi}}{2}(k+1) \right]^d & < & (k+1)^d.
\end{array}$$

We show geometric depictions of $k = 20$ multi-index sets in $d = 2$ dimensions in Figure 2. For example, we observe that $\Lambda_{\mathrm{TD}}(k) \subset \Lambda_{\mathrm{TP}}(k)$ or, in other words, that $V(\Lambda_{\mathrm{TP}}(k))$ is a larger approximation space than $V(\Lambda_{\mathrm{TD}}(k))$ and hence has a greater approximation power. However, the dimension of $V(\Lambda_{\mathrm{TP}}(k))$ is much larger than $V(\Lambda_{\mathrm{TD}}(k))$, so that more data (function evaluations) would be required to construct approximations.

This communicates the possible disadvantage of these spaces: in large dimensions polynomial subspaces V with strong approximation power may have a very large linear dimension $N = \dim V$, and would hence require a large amount of data to compute best L^2_w approximations. One must therefore balance the need for approximation accuracy against the need for computational feasibility. In practice, there are reasons to use each of these spaces; see, e.g., [54] for a motivation of Euclidean degree approximation. In this paper we are chiefly concerned with building approximations from any one of these spaces: given $V(\Lambda)$, how can we constructively compute a near-best approximation g_N to a function f?

5. Best Approximations and Least Squares.

With the notation introduced in the previous section, much of our univariate notation in section 3 can be directly applied. Let Λ be a given finite multi-index set in $\mathbb{N}_0^d$, and define $V(\Lambda)$ via (4.1); then with $N = |\Lambda|$, $\dim V = N$. Therefore, we can identify N L^2_w-orthonormal basis functions $v_1, \ldots, v_N$ satisfying (2.1). Given an L^2_w function f, we define $f_N \in V$ as the best L^2_w approximation to f from an element of V, just as in (2.3). As before, an essentially explicit way to compute f_N is furnished (2.2).

Just as in the one-dimensional case, the integrals in (2.2) usually cannot be explicitly computed and must be approximated. In contrast to the univariate case,

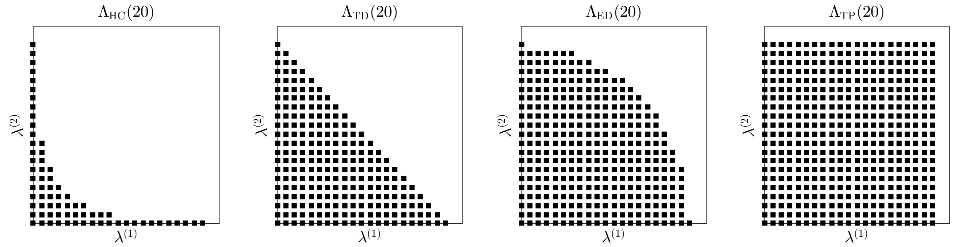
$\Lambda_{\mathrm{HC}}(20)$
$\Lambda_{\mathrm{TD}}(20)$
$\Lambda_{\mathrm{ED}}(20)$
$\Lambda_{\mathrm{TP}}(20)$
$\lambda^{(2)}$
$\lambda^{(1)}$

Fig. 2 *Visual depiction of two-dimensional multi-index sets. Left-to-right: The hyperbolic cross (HC), total degree (TD), Euclidean degree (ED), and tensor product (TP) spaces of order/degree $k = 20$. Each multi-index set Λ uniquely identifies a polynomial space $V(\Lambda)$ via the relation* (4.1).

performing discretizations in high-dimensional cases is quite difficult, for example, because there is no simple analogue of a multivariate "Gaussian" quadrature rule and many straightforward attempts at identifying highly accurate quadrature rules result in computationally infeasible constructions. A tensor-product construction illustrates the difficulty: if one forms a quadrature grid using m points per dimension, then in d dimensions this results in $M = m^d$ points. For moderate values of m and, say, $d \gtrsim 5$, the resulting computational cost (i.e., the number of times f must be evaluated) is too onerous for practical implementation.

The alternative popular strategy that we investigate in this paper is that of (randomized) discrete least squares. One reason for the popularity of this approach is that it is particularly simple to explain and implement. A discrete least-squares approximation computes the minimizer g_N of a *discrete* estimator of the norm $\|f - g_N\|$. A randomized version of this strategy chooses the discrete estimator for the norm by randomly sampling points in D. We will see that by intelligently specifying a sampling distribution for the random draw of samples, we can compute near-optimal approximations g_N with acceptable computational effort.

Suppose that $x_1, \ldots, x_M$ are M sample points in D. We will build an approximation $g_N \in V$ to f by minimizing the discrete ℓ^2 discrepancy between g_N and f on these points, i.e., we define g_N via the optimization

$$g_N \coloneqq \operatorname*{argmin}_{g \in V} \frac{1}{M} \sum_{m=1}^{M} \left(g(x_m) - f(x_m)\right)^2 . \tag{5.1}$$

In this formulation, the only information about f we require is the ensemble of data $\{f(x_m)\}_{m=1}^M$. The difference between g_N defined in this way, and f_N defined in (2.2), is in the objective function under the argmin. With f_N, the objective function is an L^2_w norm, i.e., an integral, whereas for g_N it is a discretization of this integral. To formulate the above as an algorithm, we rewrite it as a linear algebra problem. First we note that $g_N \in V$ has the form (3.3) for some coefficients c_n; we next prescribe conditions that the vector $\boldsymbol{c} = (c_1, \ldots, c_N)^T$ satisfies. Define an $M \times N$ matrix $\boldsymbol{A}$ and a vector $\boldsymbol{f} \in \mathbb{R}^M$ with entries

$$(A)_{m,n} = \frac{1}{\sqrt{M}} v_n(x_m), \qquad (f)_m = \frac{1}{\sqrt{M}} f(x_m). \tag{5.2}$$

The vector $\boldsymbol{c}$ containing expansion coefficients for g_N is defined in (5.1), which is equivalent to

$$\boldsymbol{c} \coloneqq \operatorname*{argmin}_{\boldsymbol{d} \in V} \|\boldsymbol{A}\boldsymbol{d} - \boldsymbol{f}\|_2^2 \,, \tag{5.3}$$

where $\|\cdot\|_2$ is the standard Euclidean norm on vectors. Thus, the vector $\boldsymbol{c}$ is the least-squares solution to the overdetermined linear system

$$\boldsymbol{A}\boldsymbol{c} = \boldsymbol{f}. \tag{5.4}$$

We are interested in the overdetermined case when $M \geq N$ and assume that $\boldsymbol{A}$ has full (column) rank and therefore the least-squares solution to (5.4) is unique. If $w(x) \geq \epsilon > 0$ for x in any open ball in $\mathbb{R}^d$, x_m are sampled i.i.d. from w, $M \geq N$, and V is a space of polynomials, then $\boldsymbol{A}$ has full rank with probability 1.

The approximation g_N is also defined by (5.4) through (3.3). For the purpose of analysis, the least-squares solution to (5.4) is often written as the solution to the corresponding set of *normal equations.* These equations can be derived, for example, by computing the critical point of the quadratic objective function in (5.3). If $\boldsymbol{A}$ is full rank, then the multivariate second derivative test reveals that the unique critical point is a global minimizer. The normal equations are given by

$$\boldsymbol{G}\boldsymbol{c} = \boldsymbol{g}, \qquad \boldsymbol{G} \coloneqq \boldsymbol{A}^T\boldsymbol{A} \in \mathbb{R}^{N \times N}, \qquad \boldsymbol{g} = \boldsymbol{A}^T\boldsymbol{f} \in \mathbb{R}^N, \tag{5.5}$$

which define the vector $\boldsymbol{c}$ that minimizes the ℓ^2 residual of (5.4). The system above is a square system, and hence has a unique solution when $\boldsymbol{G}$ is full rank (i.e., when $\boldsymbol{A}$ has full column rank). When $\boldsymbol{G}$ is rank-deficient, infinitely many solutions exist and any such solution achieves (5.3); this case is less practical when $M \geq N$ and we do not consider that situation in this paper. We remark that computationally one should solve (5.4) in a least-squares sense via, e.g., a QR factorization, instead of solving the normal equations system (5.5), since (5.4) is usually a much better conditioned linear system.

5.1. Random Sampling and Asymptotics. While we have completed a basic description of g_N above, it is not clear how good an approximation this provides to f. A first step in this direction can be provided by describing a particular strategy for generating the grid x_m and subsequently investigating asymptotics. Suppose that $x_1, \ldots, x_M$ are i.i.d. random draws from a random variable X with probability density w.[2] We can now motivate the large-M construction of g_N with two complementary observations.

First, consider the right-hand side argument under the argmin of (5.1). With M large, the law of large numbers implies that

$$\frac{1}{M}\sum_{m=1}^{M} \left(g(x_m) - f(x_m)\right)^2 \overset{M\uparrow\infty}{\longrightarrow} \int_D (g(x) - f(x))^2 w(x) \mathrm{d}x = \|g - f\|^2_{L^2_w},$$

so that (5.1) in the large-M limit produces the element $g_N \in V$ that is L^2_w-closest to f. This closest element is precisely the best approximation f_N in (2.3). This shows that $g_N \to f_N$ as $M \uparrow \infty$.

[2] It is more conventional to use uppercase letter notation for random variables, but we will continue to use the lowercase version x_m to denote a random variable draw.

We can likewise explore the asymptotic behavior of the entries of $\boldsymbol{G}$ and $\boldsymbol{f}$:

$$(G)_{m,n} = \frac{1}{M}\sum_{j=1}^{M} v_m(x_j)v_n(x_j) \overset{M\uparrow\infty}{\longrightarrow} \int_D v_m(x)v_n(x)w(x)\mathrm{d}x = \delta_{m,n}, \tag{5.6a}$$

$$(g)_n = \frac{1}{M}\sum_{j=1}^{M} f(x_j)v_n(x_j) \overset{M\uparrow\infty}{\longrightarrow} \int_D f(x)v_n(x)w(x)\mathrm{d}x = \widehat{f}_n. \tag{5.6b}$$

Thus, the condition (5.5) requires that c_n equals the best approximation coefficients $\widehat{f}_n$ in the large-M limit. This result is consistent with our first observation that $g_N \to f_N$ for large M. We codify this convergence as follows.

Lemma 5.1. *Let g_N be the approximation in* (3.3) *where the expansion coefficients c_n are computed as the solution to* (5.5) *or* (5.3)*. Then the following limit holds in L^2_w:*

$$\lim_{M\to\infty} g_N = f_N \ \textit{ almost surely.}$$

Proof. We will show that $c_n \to \widehat{f}_n$ almost surely, which by (3.3) and (2.2) yields the result. The strong law of large numbers establishes that the asymptotic relations (5.6) hold with probability 1. Thus, $\boldsymbol{G}$ entrywise converges to $\boldsymbol{I}$ almost surely by (5.6) as $M \uparrow \infty$. Since N is fixed and all finite-dimensional norms are equivalent, then we have the M-asymptotic results

$$\boldsymbol{G}^{-1} \to \boldsymbol{I} \ \text{ almost surely}, \qquad\qquad \boldsymbol{g} \to \widehat{\boldsymbol{f}} \ \text{ almost surely}, \tag{5.7}$$

by the continuous mapping theorem, where $\widehat{\boldsymbol{f}} = (\widehat{f}_1, \ldots, \widehat{f}_N)^T$. The limits above hold in any norm on $\mathbb{R}^N$ (vector limit) or $\mathbb{R}^{N\times N}$ (matrix limit). Now we write the normal equations (5.5) as

$$\boldsymbol{G}\boldsymbol{c} = \boldsymbol{g}, \implies \boldsymbol{c} = \boldsymbol{G}^{-1}\widehat{\boldsymbol{f}} + \boldsymbol{G}^{-1}\left(\boldsymbol{g} - \widehat{\boldsymbol{f}}\right).$$

We can now take $M \uparrow \infty$ limits of the summands above and utilize (5.7):

$$\lim_{M\to\infty} \boldsymbol{G}^{-1}\widehat{\boldsymbol{f}} = \boldsymbol{I}\widehat{\boldsymbol{f}} = \widehat{\boldsymbol{f}},$$
$$\lim_{M\to\infty} \boldsymbol{G}^{-1}\left(\boldsymbol{g} - \widehat{\boldsymbol{f}}\right) = \boldsymbol{I}\boldsymbol{0} = \boldsymbol{0}.$$

Both the above limits hold almost surely in any norm on $\mathbb{R}^N$. Therefore, we have established that $\boldsymbol{c} \to \widehat{\boldsymbol{f}}$ with probability 1. □

The result above establishes that a discrete least-squares approximation g_N built from random sampling M-*asymptotically* achieves the best possible approximation. However, this does not establish quantitative error in the preasymptotic regime. Recall our definition of η_N in (3.4), and the fact that in the one-dimensional example shown in Figure 1, we achieved $\eta_N \approx 1$, indicating that the function g_N is comparable in L^2_w-approximation quality to the best, usually uncomputable, approximation f_N. Since g_N is uniquely identified by N coefficients (the c_n), then one hopes for an outcome similar to the one-dimensional case with Gaussian quadrature illustrated in Figure 1: approximately $M \sim N$ samples can allow us to compute a g_N so that $\eta_N \sim 1$. It turns out that, in practice, we frequently cannot achieve this as above by sampling x_m i.i.d. from the density w.

EXAMPLE 5.2. *Consider the Gaussian density case* $w(x) = \exp(-\|x\|^2)/\pi$ *for* $x \in D = \mathbb{R}^2$. *Given a positive integer* k, *we define* $\Lambda = \Lambda_{\mathrm{TD}}(k)$, *which identifies* $V = V(\Lambda)$ *through* (4.1). *Consider the test function* $f(x)$ *defined as*

$$f(x) = B(\|x/4 - (0.2, -0.1)\|_2),$$

where $B(x)$ *is the univariate bump function in* (3.5). *We will test approximations of this function from the space* V *for* $k = 1, \ldots, 25$. *Note that* $N = \dim V = \binom{k+2}{k} = (k+1)(k+2)/2$ *in this case, so that* N *grows quadratically with* k. *We compute least-squares approximations* g_N *from function samples, where the samples* x_m *are taken as i.i.d. samples from the density* w. *We take* $M = 10N$ *samples to construct* g_N, *which is quite a considerable oversampling. Figure* 3 *(left) shows the results of this experiment: the least-squares estimator* g_N *is extremely inaccurate, but the best approximation* f_N *should be a reasonable approximation. To study how* g_N *behaves as a function of the sample count* M, *we next fix* $k = 20$, *which also fixes* $N = 231$, *and now test various sample size ranges from* $M = N$ *to* $M = 10N$. *The results are shown in Figure* 3 *(center) and illustrate that* g_N *computed in this way converges extremely slowly, i.e.,* η_N *is quite large even for* $M = 10N$. *We will return to this example in section* 7 *after introducing a particular weighted least-squares procedure using samples from the induced distribution. As shown in Figure* 3 *(right) this weighted least-squares procedure substantially improves stability and accuracy of the approximation* g_N.

For completeness, the rightmost pane of Figure 3 illustrates the more positive result of performing discrete least squares using a particular biased sampling technique that is the main focus of this paper and serves to ameliorate the bad behavior observed in the left and center panes. The details of this sampling technique are explained in what follows.

5.2. Biased Sampling. In the previous section we formed a least-squares approximation g_N by taking $x_1, \ldots, x_M$ defining the normal equations (5.5) to be i.i.d. samples from a random variable with density w. In this section, we will generalize this idea to the case of *biased* sampling in a weighted least-squares formulation. To be precise, let $q(x)$ denote any positive L^2_w function with unit L^2_w norm, i.e., $\|q\| = 1$; this choice implies that

$$\int_D \rho(x)\mathrm{d}x = 1, \qquad \rho(x) := q^2(x)w(x), \tag{5.8}$$

so that $\rho(x)$ is another probability density over D. Note that $q(x) \equiv 1$ is one such choice, since w is itself a probability density function. We will consider forming a $1/q^2$-weighted least-squares approximation by taking samples from a random variable with density $\rho = q^2 w$.

Let $x_1, \ldots, x_M$ denote i.i.d. samples drawn from a random variable whose density is ρ. We now generalize the definition of the matrices $\boldsymbol{A}$ and the vector $\boldsymbol{f}$ introduced in (5.2) by building in dependence on q:

$$(A)_{m,n} = \frac{1}{\sqrt{Mq^2(x_m)}} v_n(x_m), \qquad (f)_m = \frac{1}{\sqrt{Mq^2(x_m)}} f(x_m). \tag{5.9}$$

We emphasize that if $q(x) \equiv 1$, then the definitions above revert to those in (5.2). Note that we have now *weighted* each sample, so that the least-squares solution to the overdetermined system

$$\boldsymbol{A}\boldsymbol{c} = \boldsymbol{f}$$

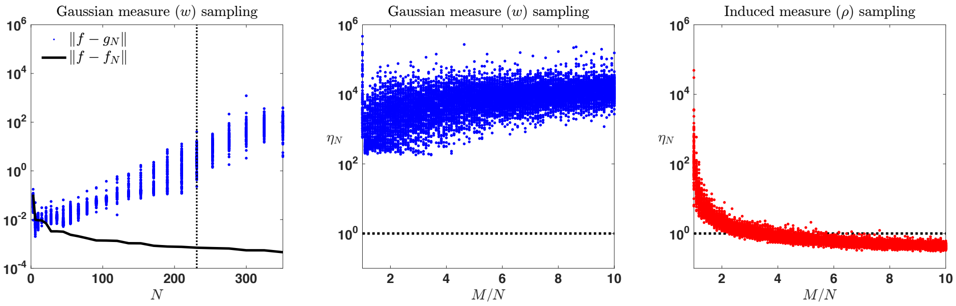

Fig. 3 *Companion results for Example 5.2. Left: Comparison between the best approximation error $\|f - f_N\|$ and a "standard" discrete least-squares approximation error $\|f - g_N\|$, where the samples are drawn i.i.d. from w. We use $M = 10N$ samples to generate g_N. Since the procedure for constructing g_N is random, 100 trials are performed for each value of N, and the results from the collection of trials are presented as a scatter plot. The vertical black dotted line indicates the value of N used in the center and right plots. Center: Fixing a degree $k = 20$ approximation ($N = 231$) shows values of the relative error metric η_N defined in (3.4) for the "standard" discrete least-squares approximation as a function of the number of samples M used in the construction. The large values of η_N show that this method of construction produces poor approximations. Right: For the same setup as the center pane, this shows construction of an approximation g_N using induced distribution sampling, which is the main topic of this paper, where samples are drawn i.i.d. from ρ introduced in section 7. This latter approach yields values $\eta_N \sim 1$ with moderate dependence of M on N, motivating its use in practice.*

is a *weighted* least-squares formulation, i.e., we are solving (5.3) with our new definitions of $\boldsymbol{A}$ and $\boldsymbol{f}$ and forming the approximation g_N via the optimization

$$g_N := \operatorname*{argmin}_{g \in V} \frac{1}{M} \sum_{m=1}^{M} \frac{(g(x_m) - f(x_m))^2}{q(x_m)^2}.$$

The resulting coefficient vector $\boldsymbol{c}$ depends on the choice of q, i.e., depends on the choice of our weighting. Note that, with our new definitions of $\boldsymbol{A}$ and $\boldsymbol{f}$, the normal equations are formulated precisely as in (5.5).

At face value, it does not seem that we have accomplished anything. Indeed, since we have sampled from the density $\rho = q^2 w$ and weighted by $1/q^2$, then in the large-M limit, this change of variable results in asymptotic properties that are independent of q:

$$\frac{1}{M} \sum_{m=1}^{M} \frac{(g(x_m) - f(x_m))^2}{q(x_m)^2} \xrightarrow{M\uparrow\infty} \int_D \frac{(g(x) - f(x))^2}{q^2(x)} q^2(x) w(x) \mathrm{d}x = \|g - f\|^2_{L^2_w}.$$

In addition, we have results analogous to (5.6):

$$(G)_{m,n} = \frac{1}{M} \sum_{j=1}^{M} \frac{v_m(x_j) v_n(x_j)}{q^2(x_j)} \xrightarrow{M\uparrow\infty} \int_D \frac{v_m(x) v_n(x)}{q^2(x)} q^2(x) w(x) \mathrm{d}x = \delta_{m,n},$$

$$(g)_n = \frac{1}{M} \sum_{j=1}^{M} \frac{f(x_j) v_n(x_j)}{q^2(x_j)} \xrightarrow{M\uparrow\infty} \int_D f(x) \frac{v_n(x)}{q^2(x)} q^2(x) w(x) \mathrm{d}x = \widehat{f}_n.$$

Indeed, all we have done is change our discrete estimation of the integrals on the right-hand side by changing the sampling distribution. Such an idea is hardly new and is the basis for biased or importance sampling strategies. Just as in those contexts, we can choose q to mitigate the instabilities observed when constructing g_N in Figure 3. The fundamental analysis that demonstrates the optimal choice of q centers around the normal equations.

Due to the M-asymptotic limits above, the M-asymptotic behavior of g_N is just as in Lemma 5.1: we have $g_N \to f_N$ with probability 1 as $M \to \infty$. However, a more interesting topic is what happens in the preasymptotic regime.

6. Preasymptotic Analysis of the Normal Equations. A qualitative understanding of when the approximation g_N is close to the best approximation f_N can be given by inspection of the normal equations defined in (5.5) (where again the matrices $\boldsymbol{A}$ and $\boldsymbol{f}$ are defined in (5.9) with an as yet unspecified function q). We reiterate that we have

$$\boldsymbol{G} \overset{M\uparrow\infty}{\longrightarrow} \boldsymbol{I}, \qquad\qquad \boldsymbol{g} \overset{M\uparrow\infty}{\longrightarrow} \left(\widehat{f}_1, \ldots, \widehat{f}_N\right)^T,$$

where $\boldsymbol{I}$ is the $N \times N$ identity matrix. Thus, when $\boldsymbol{G}$ is "close" to $\boldsymbol{I}$ and $\boldsymbol{g}$ has entries "close" to $\widehat{f}_n$, the g_N will be "close" to f_N (cf. the proof of Lemma 5.1). In this section we will concentrate on providing analysis that quantifies the proximity of $\boldsymbol{G}$ to $\boldsymbol{I}$, indicating that the linear system of normal equations (5.5) is well-conditioned and that the least-squares problem is therefore stable. A full, rigorous analysis that uses this to quantify the proximity of g_N to f is too technical for this paper, but is not needed to communicate the main ideas about how one can choose q to prevent instabilities in least-squares computations.

We present the main result we need, which is a specialization of Theorem 1 in [14], followed by a brief proof.

Theorem 6.1 ([14]). *Define $\boldsymbol{G}$ as in (5.5) with $\boldsymbol{A}$ defined in (5.9). If, for some $r > 0$, the number of samples M satisfies*

$$\frac{M}{\log M} \geq C(r+1) \sup_{x \in D} \sum_{n=1}^{N} \left(\frac{v_n(x)}{q(x)}\right)^2, \tag{6.1}$$

where $C = 2/\log(27/8e) \approx 9.24$, then with probability at least $1 - 2M^{-r}$ we have

$$\|\boldsymbol{G} - \boldsymbol{I}\|_2 \leq \frac{1}{2},$$

where $\|\cdot\|_2$ is the induced ℓ^2 norm on matrices.

Proof. The entries of $\boldsymbol{G}$ defined by (5.5) have the expression

$$(G)_{j,k} = \sum_{m=1}^{M} \frac{1}{Mq^2(x_m)} v_j(x_m) v_k(x_m), \qquad 1 \leq j, k \leq N,$$

and hence we have the matrix equality

$$\boldsymbol{G} = \sum_{m=1}^{M} \boldsymbol{V}_m, \qquad (V_m)_{j,k} = \frac{1}{Mq^2(x_m)} v_j(x_m) v_k(x_m), \qquad 1 \leq j, k \leq N.$$

Since $\boldsymbol{V}_m$ depends only on x_m, and since x_m are drawn as i.i.d. samples (from the density $\rho = q^2 w$), then $\boldsymbol{V}_m$ are i.i.d. matrices, and in particular are clearly symmetric.

Furthermore, each $\boldsymbol{V}_m$ is a nonnegative matrix, here meaning all eigenvalues are nonnegative. This is easily verified from the above expression by noting that $\boldsymbol{V}_m$ is the rank-one matrix

$$\boldsymbol{V}_m = \boldsymbol{v}_m \boldsymbol{v}_m^T, \qquad \boldsymbol{v}_m^T = \frac{1}{\sqrt{M} q(x_m)} \left(v_1(x_m), \dots, v_N(x_m)\right)^T.$$

Thus, we have shown explicitly that $\boldsymbol{V}_m$ has one eigenvalue equal to $\|\boldsymbol{v}_m\|^2 \geq 0$ and that its remaining eigenvalues are 0; thus, $\boldsymbol{V}_m$ is symmetric and positive (semi)definite.

One final piece we require is the simple computation

$$\mathbb{E}(V_m)_{j,k} = \int_D \frac{1}{M} \frac{v_j(x) v_k(x)}{q^2(x)} q^2(x) w(x) \mathrm{d}x = \frac{1}{M} \delta_{j,k},$$

where the expectation is taken over the density $\rho = q^2 w$, corresponding to the random draw of x_m. This in turn implies that $\mathbb{E}\boldsymbol{G} = \sum_{m=1}^M \mathbb{E}\boldsymbol{V}_m = \boldsymbol{I}$.

In order to understand how far $\boldsymbol{G}$ deviates from its expectation, the identity matrix, we employ a matrix Chernoff bound for sums of i.i.d. positive semidefinite matrices [56]. Let $\lambda_{\min}(\boldsymbol{M})$ and $\lambda_{\max}(\boldsymbol{M})$ denote the minimum and maximum eigenvalues, respectively, of a symmetric matrix $\boldsymbol{M}$. The Chernoff bound we use will specify how far the spectrum of $\boldsymbol{G}$ deviates from the spectrum of its expectation $\mathbb{E}\boldsymbol{G}$. To this end, let

$$\tau_{\min} \coloneqq \lambda_{\min}\left(\mathbb{E}\boldsymbol{G}\right) = 1, \qquad \tau_{\max} \coloneqq \lambda_{\max}\left(\mathbb{E}\boldsymbol{G}\right) = 1.$$

Finally, let Q denote a bound on the spectral norm of the summand $\boldsymbol{V}_m$:

(6.2)
$$\lambda_{\max}(\boldsymbol{V}_m) = \|\boldsymbol{V}_m\|_2 = \|\boldsymbol{v}_m\|^2 = \frac{1}{M} \sum_{n=1}^N \left(\frac{v_n(x_m)}{q(x_m)}\right)^2 \leq \frac{1}{M} \sup_{x \in D} \sum_{n=1}^N \left(\frac{v_n(x)}{q(x)}\right)^2 =: Q.$$

We will be interested in the following events:

$$E_{\min} \coloneqq \left\{\lambda_{\min}\left(\boldsymbol{G}\right) \leq \frac{1}{2}\tau_{\min}\right\}, \qquad E_{\max} \coloneqq \left\{\lambda_{\max}\left(\boldsymbol{G}\right) \geq \frac{3}{2}\tau_{\max}\right\},$$
$$E \coloneqq E_{\min} \bigcup E_{\max} = \left\{\|\boldsymbol{G} - \boldsymbol{I}\|_2 > \frac{1}{2}\right\}.$$

In particular, E is an event corresponding to a relatively ill-conditioned set of normal equations, so that we hope to minimize the probability of E. The matrix Chernoff bound we employ, discussed in [56], gives bounds on the probability that the spectrum of $\boldsymbol{G}$ deviates substantially from that of $\mathbb{E}\boldsymbol{G}$:

$$\Pr\left[E_{\min}\right] \leq N \left(\frac{2}{e}\right)^{\tau_{\min}/2Q}, \qquad \Pr\left[E_{\max}\right] \leq N \left(\frac{8e}{27}\right)^{\tau_{\max}/2Q}.$$

The expression simplifies a bit by noting that $\tau_{\min} = \tau_{\max} = 1$. We can now use the union bound

$$\Pr\left[E_{\min} \bigcup E_{\max}\right] \leq \Pr\left[E_{\min}\right] + \Pr\left[E_{\max}\right],$$

and use the fact that $2/e < 8e/27$ to conclude that

$$\Pr[E] \leq 2N\left(\frac{8e}{27}\right)^{1/2Q} = 2N\exp\left(\frac{-1}{2Q}\log(27/8e)\right).$$

Therefore, if we choose a number of samples M such that

$$\frac{\log(27/8e)}{2Q} \geq (r+1)\log M, \tag{6.3}$$

then this guarantees

$$\Pr[E] \leq 2N\exp\left(-(r+1)\log M\right) = 2NM^{-r}M^{-1} \leq 2M^{-r}.$$

The condition (6.3) is equivalent to (6.1), proving the result. □

7. Optimal Sampling and Least Squares.

The result of Theorem 6.1 allows us to devise a sampling strategy that yields stable discrete least-squares problems in high dimensions. The main strength of this theorem is (6.1), specifying how large M should be to ensure stability with high probability. The only portion of the condition (6.1) that depends on the density w, the domain D, or the dimension d is the quantity

$$\sup_{x\in D}\sum_{n=1}^{N}\left(\frac{v_n(x)}{q(x)}\right)^2.$$

Therefore, we seek to choose q so that this quantity is minimized. We note that in general the above expression can be no smaller than N:

$$\sup_{x\in D}\sum_{n=1}^{N}\left(\frac{v_n(x)}{q(x)}\right)^2 \geq \int_D \sum_{n=1}^{N}\left(\frac{v_n(x)}{q(x)}\right)^2 \rho(x)\mathrm{d}x = \sum_{n=1}^{N}\int_D v_n^2(x)w(x)\mathrm{d}x = N,$$

where the inequality uses the fact that $\rho(x) = q^2(x)w(x)$ is a probability density on D, and the last equality is true since $\{v_n\}_{n=1}^N$ is an L^2_w-orthonormal basis. The authors in [16] observe that one can actually achieve this lower bound exactly by choosing $q^2(x)$ as

$$q^2(x) = \frac{1}{N}\sum_{n=1}^{N}v_n^2(x) \quad \Longrightarrow \quad \sup_{x\in D}\sum_{n=1}^{N}\left(\frac{v_n(x)}{q(x)}\right)^2 = N. \tag{7.1}$$

This in turn implies that if x_m are i.i.d. samples from the density $\rho = q^2 w$, where q is defined as in (7.1), then the sampling criterion (6.1) that ensures a stable least-squares formulation with high probability is

$$\frac{M}{\log M} \geq C(r+1)N, \qquad\qquad C = \frac{2}{\log(27/8e)}, \tag{7.2}$$

which, apart from logarithmic factors, is optimal in terms of the dependence of M on N. Note that in practical scenarios the user-defined approximation space V will depend on the dimension d (cf. the discussion in section 4 showing that many standard polynomial approximation spaces V have dimension that depends on d). However, the least-squares stability criterion (7.2) depends *only* on $\dim V = N$ and is *independent* of the dimension d, the domain D, the density w, and even the particular choice of N-dimensional subspace V. The tradeoff is that we must sample from the rather nonstandard density ρ, which *does* depend on (V, w, D).

Definition 7.1. *Let D, w, and V be given so that $V \subset L^2_w(D)$. The (V, w, D)-induced distribution corresponds to the following probability density on D:*

$$\rho(x) = \frac{1}{N} \sum_{n=1}^{N} v_n^2(x) w(x), \tag{7.3}$$

where $\{v_n\}_{n=1}^N$ is any L^2_w-orthonormal basis for V.

The terminology *induced* is borrowed from similar terminology in orthogonal polynomials [23]. Comparing (7.3) with (5.8), we see that the definition above corresponds to the choice

$$q^2(x) = \frac{1}{N} \sum_{n=1}^{N} v_n^2(x).$$

At first glance the density ρ in (7.3) appears to depend on the choice of orthonormal basis element v_n. However, any unitary transformation of $\{v_n\}_{n=1}^N$ leaves the quadratic form $\sum_n v_n^2$ unchanged, so that in reality the function defined in (7.3) is a property of the subspace V and not of the chosen basis.

Our remedy for the unstable observations in Figure 3 is thus as follows: sample x_m i.i.d. from the induced density ρ and perform a weighted least-squares approximation, where the weights are the explicit quantities $1/q(x_m)^2 = w(x_m)/\rho(x_m)$. If we perform this type of biased sampling and weighted least-squares procedure, we obtain the results shown in the rightmost pane of Figure 3, illustrating that we have substantially ameliorated the instability.

7.1. Convergence. Until now we have only established a preasymptotic characterization of $\boldsymbol{G} - \boldsymbol{I}$, but have not discussed how this can be translated into g_N converging to f_N. A full discussion and proof are outside the scope of this paper, but we provide the following result from [16] that establishes one such convergence result.

To proceed we need to introduce an additional assumption, namely, that f is bounded:

$$\sup_{x \in D} |f(x)| = L < \infty. \tag{7.4}$$

Furthermore, we will need a truncation operator T_L, defined as

$$T_L(y) \coloneqq \operatorname{sign}(y) \min\{|y|, L\}.$$

In particular, we can use this to define $T_L \circ g_N$, which is an L-truncated version of g_N and is the focus of the following result.

Theorem 7.2 ([16]). *Let (V, w, D) be given and fixed, and assume $f \in L^2_w$ satisfies* (7.4). *Assume x_m, $m = 1, \ldots, M$, are i.i.d. samples from the induced density $\rho = wq^2$ in* (7.3), *and assume that for some $r > 0$ the number of samples M satisfies* (6.1). *Let g_N be constructed with a weighted least-squares approach in which the coefficients $\boldsymbol{c}$ are given by the solution to* (5.5), *where $\boldsymbol{A}$ is defined in* (5.9). *Noting that g_N is a random function, we have*

$$\mathbb{E} \, \|f - T_L \circ g_N\|^2 \leq \left(1 + \frac{4}{C(1+r)\log M}\right) \|f - f_N\|^2 + 8L^2 M^{-r},$$

where C is the constant in (6.1).

We refer the reader to [16] for the proof, and note that it hinges on first showing the proximity of $\boldsymbol{G}$ to $\boldsymbol{I}$, which we established earlier. The result above establishes that, in expectation, a truncated version of the least-squares approximation g_N commits an error comparable to the best possible error committed by f_N. One can also prove statements in high probability about the nontruncated approximation g_N.

8. Asymptotic Optimal Sampling.

We have seen that a sampling i.i.d. from the nonstandard induced density ρ given by (7.3) can substantially improve least-squares approximations compared to to sampling i.i.d. from the original density w. A natural desire is to seek to understand the density ρ, and in the context of polynomial approximation a first step to accomplishing this is studying how ρ behaves as the polynomial degree tends to infinity.

The large-degree limit of induced distributions is known in the one-dimensional case. With V the space of polynomials of degree $N-1$ and lower, behavior of ρ as $N \uparrow \infty$ is known in some generality through potential theory. For example, with $w(x)$ uniform on $[-1,1]$, the large-N limit of ρ is

$$\lim_{N\to\infty} \rho(x) = \rho_\infty(x) := \frac{1}{\pi\sqrt{1-x^2}},$$

where the limit is true in the weak sense. This shows that sampling according to the "Chebyshev" density is the large-N optimal strategy [45]; such a result is consistent with the classical numerical analysis knowledge that polynomial approximation on bounded intervals with point evaluations is best accomplished with a Chebyshev-distributed grid instead of a uniform grid. Similar results about ρ_∞ are known for a variety of univariate densities w. These limiting densities are summarized in [44] with the references therein being the appropriate historical and seminal sources. Figure 4 compares the "standard" density w, the induced density ρ, and the N-asymptotic induced density ρ_∞ for three different choices of univariate w and with V the space of polynomials up to degree $N-1=19$. One can observe that while ρ and ρ_∞ are visually close, w and ρ can be substantially different. In particular, ρ can dictate substantial sampling in regions of D where w would require very few samples.

For the multivariate case, much more is unknown, and in many cases we currently have only conjectures about such limiting densities. A similar experiment in the multivariate case is complicated by the fact that many polynomial spaces V can be defined associated to a particular "degree" k, since there are different ways of defining degree-k multi-index sets Λ. For example, with Λ_{TD}, Λ_{ED}, Λ_{TP}, and Λ_{HC} as defined in (4.2), we have enumerated four different ways to choose a space of "degree" k, resulting in four different choices of a "degree k" induced distribution for degree k:

$$V(\Lambda_{\mathrm{TP}}(k)) \to \rho_{\mathrm{TP}}, \quad V(\Lambda_{\mathrm{ED}}(k)) \to \rho_{\mathrm{ED}}, \quad V(\Lambda_{\mathrm{TD}}(k)) \to \rho_{\mathrm{TD}}, \quad V(\Lambda_{\mathrm{HC}}(k)) \to \rho_{\mathrm{HC}}.$$

Large-k asymptotics for these cases is largely understudied, but the total degree case has received some attention with some existing conjectures for total degree asymptotics [45]. Consider $D=[-1,1]^d$ with w uniform on D, and let $V=V(\Lambda_{\mathrm{TD}}(k))$. Then the large-$k$ asymptotic density is known as the tensorized Chebyshev density, i.e.,

$$\lim_{N\to\infty} \rho(x) = \frac{1}{\pi^d \prod_{j=1}^d \sqrt{1-\left(x^{(j)}\right)^2}}.$$

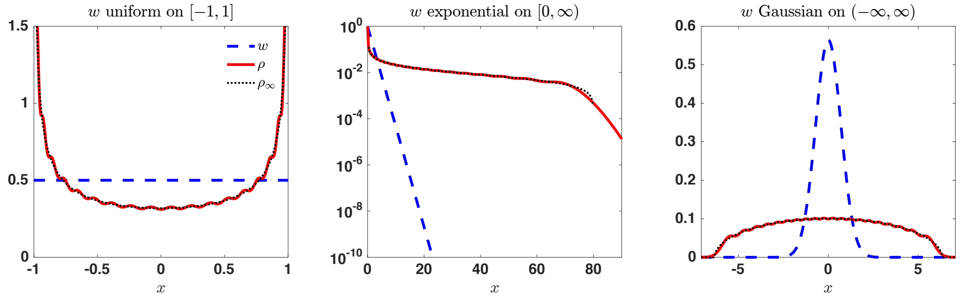

Fig. 4 *Plots of induced distribution probability density functions ρ associated to various densities w for $N = 20$, where V is the space of polynomials of degree $N-1$ and lower. Also shown are the N-asymptotic limits ρ_∞. Left: $w(x)$ uniform on $[-1,1]^2$. Middle: w exponential on the positive real line. Right: w Gaussian on the real line.*

Notice that sampling with the Chebyshev density is straightforward. On the other hand, consider $D = R^d$ with w the Gaussian density as in Example 5.2. The authors in [45] conjecture that the asymptotic density is of the form

$$\lim_{k\to\infty} \rho(x/\sqrt{k}) = C\left[2 - \|x\|_2^2\right]^{d/2}, \tag{8.1}$$

where C is a normalizing constant so that the limit is a probability density. Notice that under the limit we require the input to ρ to be scaled by $1/\sqrt{k}$. Efficient sampling schemes with the above density can be found in [45]. These examples demonstrate that k-asymptotic sampling strategies can be characterized, but that they are sometimes not obvious.

The characterization of k-asymptotic univariate induced densities hinges on the notion of equilibrium measures from weighted potential theory [49]. A similar notion of equilibrium measures in the multivariate case can be formulated through weighted pluripotential theory [35, 5]. The authors in [45] use this connection to formulate the asymptotic results above and to propose a least-squares sampling strategy based on equilibrium measures. While the strategy is only k-asymptotically optimal, it utilizes sampling from only standard distributions for a very wide variety of densities w. This may be advantageous in applications, such as in adaptive approximation schemes where the polynomial space is adaptively constructed. Another application is in the so-called data-driven approach, where w is unknown except for perhaps its support, and moments are approximated from an available database of samples from w [48, 24].

8.1. Sampling Scheme for the Induced Distribution. A final point that merits discussion in this section is the task of sampling from ρ in (7.3). While sampling from *general* multivariate probability densities is computationally onerous, the formula (7.3) is an additive mixture of tensor-product densities, and so can easily be sampled with linear complexity in the dimension d. More discussion on this topic is provided in [16, 42], with software implementing this sampling provided in [41].

Some induced densities are compared in two dimensions in Figure 5 for w a Gaussian density on $\mathbb{R}^2$ for $k = 8$. Again we see that w differs substantially from each induced distribution, but in addition we see that, e.g., the Euclidean degree and hyperbolic cross induced distributions also differ substantially.

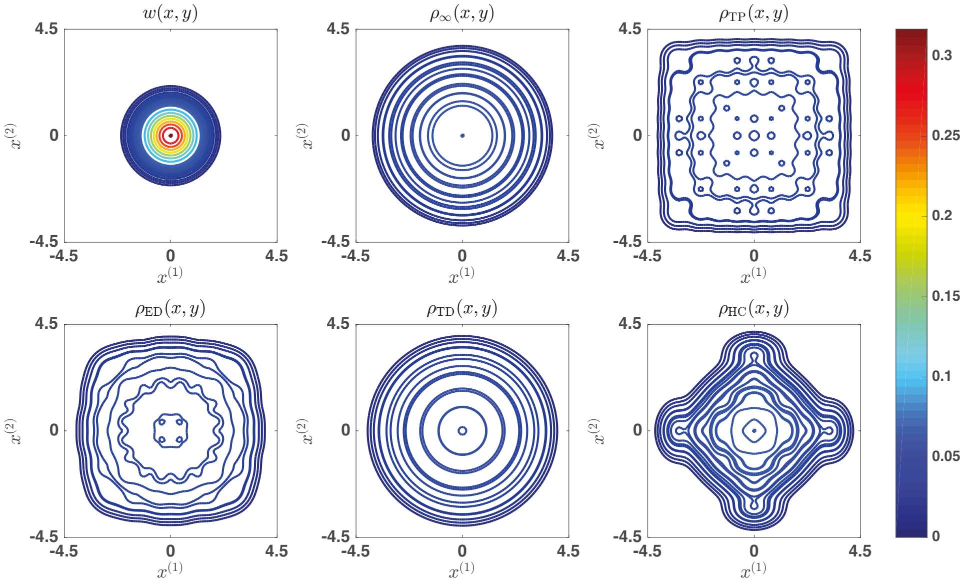

Fig. 5 *Contour plots of induced distribution probability density functions ρ associated to the density $w(x) = \exp(-\|x\|_2^2)/\pi$ for $x \in \mathbb{R}^2$, with V a space of degree $k = 8$. We show results for the induced densities associated to total degree (TD), Euclidean degree (ED), tensor product (TP), and hyperbolic cross (HC) spaces. Also shown is a conjectured density ρ_∞ for the large-k total-degree limiting density.*

EXAMPLE 8.1. *We consider approximations in the space L^2_w with $w = \exp(-\|x\|_2^2)/\pi^{d/2}$ for $x \in \mathbb{R}^d$. Our test function will itself be a Gaussian bump,*

$$f(x) = \prod_{j=1}^{d} \exp\left(\left[x^{(j)}\right]^2/j\right), \qquad x = \left(x^{(1)}, x^{(2)}, \ldots, x^{(d)}\right).$$

We have intentionally chosen a test function f that is of product form so that we can easily compute best approximations f_N in high dimensions. We will test approximations from the space $V(\Lambda)$ with $\Lambda = \Lambda_{\mathrm{HC}}(k)$. We investigate three different pairs (d,k):

$$(d,k) = (4,20),\ (8,10),\ (20,5).$$

For each test, we compute the relative error metrics η_N defined in (3.4) *using a least-squares approximation g_N built* (a) *from "standard" i.i.d. samples from w using unweighted least squares, and* (b) *from induced distribution samples from ρ using weighted least squares. The results are shown in Figure* 6*, illustrating that induced distribution sampling outperforms standard sampling consistently, but the advantage diminishes for large dimension. The reason for this diminishing advantage is that the space $V(\Lambda)$ for smaller values of k (i.e., larger values of d) has low-degree polynomials, and in this case the induced distribution density ρ defined in* (7.3) *is close to w. Nevertheless, one observes that the qualitative accuracy behavior of g_N using weighted least squares with the induced distribution is essentially unchanged as the dimension d increases, which is the expected behavior given Theorems* 6.1 *and* 7.2.

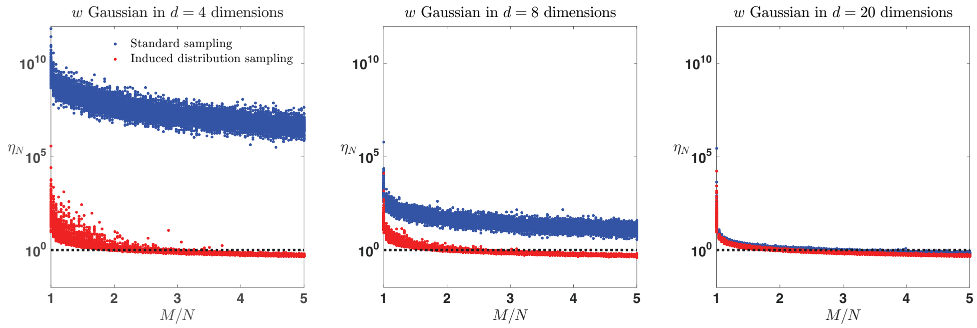

Fig. 6 *Figure associated to Example* 8.1*. Relative errors η_N when the least-squares approximation g_N is built using samples from w (blue dots) versus from the induced distribution ρ (red dots). Since the procedure is randomized,* 100 *trials are performed for each value of M, and the results are presented as a scatter plot. The approximation space V corresponds to $\Lambda_{\mathrm{HC}}(k)$, and w is the Gaussian density on $\mathbb{R}^d$. Left:* $(d, k) = (4, 20)$*. Middle:* $(d, k) = (8, 10)$*. Right:* $(d, k) = (20, 5)$*.*

Remark 8.2. We close this section by remarking that throughout this paper we have only considered random sampling schemes. However, quasi-random or deterministic sampling schemes are also interesting and useful [61, 39, 25].

9. Example. We now present a thermal diffusion problem to show the applications of the least-squares approaches in parametric uncertainty quantification. The problem is defined in the unit square following [2]:

$$\begin{cases} -\nabla \cdot (a(y,x)\nabla u(y,x)) = S(y,x) & \text{for}(y,x) \in \ \ \Omega \times D, \\ u(y,x) = 0 & \text{for}(y,x) \in \partial\Omega \times D. \end{cases} \tag{9.1}$$

The domain Ω is a two-dimensional square domain depicted in Figure 7, left. The conductivity coefficient a depends on a finite number of random variables (parameters) $x \in D$. We consider a test case with forcing term $S(y, x) = 100\chi_F(y)$, where $\chi_F(y)$ is the indicator function of the domain $F \subset \Omega$, a square subdomain with side length equal to 0.2 centered in the domain as shown in Figure 7. The material features four circular inclusions with radius $r = 0.13$ and symmetrically distributed with respect to the center of the square, each with a different conductivity value, i.e., we take the conductivity coefficient as $a(y, x) = 1 + \sum_{i=1}^{4} x^{(i)}\chi^i(y)$, where $\chi^i(y)$ is the indicator function for each circle shown in Figure 7. Thus, $x \in D \subset \mathbb{R}^4$. The quantity of interest that we approximate is defined as

$$f(x) := u\,(y, x)\,, \qquad\qquad y = (0.25, 0.375) \in \Omega.$$

We interpret the parameters $x^{(i)}$ as i.i.d. random variables, each having a uniform distribution, $x^{(i)} \sim \mathbb{U}(-0.99, 0.2)$. We are therefore interested in constructing an approximation to the map $x \mapsto f(x)$ in the L^2_w norm, where w is the joint probability density function of x.

Since w is uniform on a hypercube, the associated polynomial basis functions are tensorized Legendre polynomials. For the spatial (y) discretization, we use the finite element method with P_1 elements. The procedure for evaluating $f(x)$ then proceeds as $x \mapsto u(y, x) \mapsto u((0.25, 0.375), x)$. Hence, each evaluation of f requires a solution

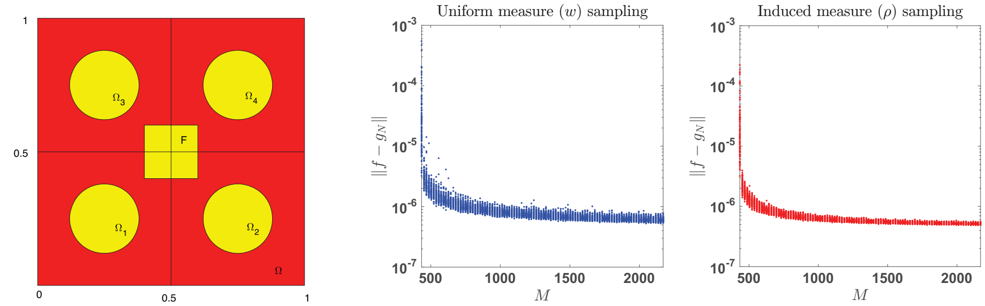

Fig. 7 *Figure associated to PDE example* (9.1). *Left: Geometry for test case. Middle and right: Discrete ℓ_2 error against sampling number obtained from uniform sampling and weighted least-squares sampling from the induced distribution, respectively.* 100 *trials are performed for each value of M. The approximation space V corresponds to $\Lambda_{\mathrm{HC}}(k)$ with $(d,k) = (4,25)$.*

to a computationally involved finite element problem, motivating the need to design algorithms that evaluate f as few times as possible.

To evaluate error in the procedure, we collect $Q = 10{,}000$ samples $\{f(z_j)\}_{j=1}^{Q}$ to define the following discrete ℓ_2 error:

$$\|f - g_N\| \approx \left(\frac{1}{M} \sum_{j=1}^{M} (g_N(x_j) - f(x_j))^2 \right)^{1/2} .$$

Here $g_N(x_j)$ represents the least-squares solution computed via either uniform density sampling or the induced distribution sampling. The approximation space V is chosen as $V = V(\Lambda_{\mathrm{HC}}(25))$. The numerical results are shown in the center and right plots of Figure 7.

We observe that while weighted least-squares sampling from the induced distribution produces results that are perhaps a bit better than the standard sampling results, the difference is not substantial. This observation mirrors the lesson in Example 8.1 that in some cases standard sampling from w is competitive with induced distribution sampling from ρ, but the induced distribution sampling is consistently among the best, and is backed by theory such as Theorems 6.1 and 7.2 requiring a minimal sample count.

10. Conclusion.

The construction of least-squares polynomial approximations to functions in multiple dimensions is a useful computational approximation strategy. When approximating functions of high-dimensional inputs x, the technique of randomly choosing sample points is preferred to a grid formed from a computationally infeasible tessellation of high-dimensional space. When approximating functions in an L^2_w sense, it makes sense to generate random samples from the density w. However, a better strategy is to sample from a biased density (weighting accordingly to account for the bias).

We have shown how random matrix concentration arguments can be used to derive a sample count condition that guarantees stability of the discrete least-squares problem. This condition gives rise to an optimal sampling density, the *induced distribution*, and sampling data from this density requires only log-linear dependence of the number of samples M on the dimension of the approximation space N; see

(7.2). In particular, we have demonstrated that sampling from w can often be significantly suboptimal in practice, resulting in ill-behaved least-squares approximations, but that a weighted least-squares approximation using the induced distribution provides dimension-independent behavior of least-squares stability and accuracy.

11. Extensions and New Directions.

In the context of the literature, our investigations have been relatively limited: we have considered computing least-squares approximations on a discrete grid that is generated by i.i.d. samples. There are numerous alternatives and extensions that have been investigated in computing approximations in high dimensions:

- We have investigated the so-called *noiseless* case, assuming that the data vectors $\boldsymbol{f}$ and/or $\boldsymbol{g}$ are exact function evaluations. One can successfully show that the least-squares procedure is stable with respect to noisy data [14, 16].
- Much of our discussion has centered around polynomial approximations, but all results outside of section 8 can be applied to general spaces of functions in L^2_w, not just polynomials.
- One could consider interpolative reconstructions with a variety of meshes and polynomial spaces [7, 6, 46, 47], but analysis of convergence in this case is much more difficult.
- When the number of samples M is very large, the requisite numerical linear algebra in this paper may be computationally infeasible. In this case one can investigate randomized linear algebra algorithms to compute solutions to least-squares problems [52, 50].
- Knowledge of the appropriate approximation space V is required to perform least squares as described in this paper. However, a more promising approach is to learn the appropriate subspace V from given data. Much work has been devoted to such adaptive approximation techniques [38, 17, 12].
- One may use quasi-Monte Carlo low-discrepancy point sets [37, 18], which have been investigated in [39].
- Alternative discrete or quasi-discrete point sets can produce good least-squares approximations. These sets are sometimes generated deterministically [10, 61] or through an optimization process [25, 51, 29].
- Large parametric dimensions d often yield high approximation space dimensions N, and thus require a large amount of data. When data is scarce, $M \ll N$, one can seek approximations whose expansion coefficients are sparse or compressible. The techniques to compute such approximations hinge on theory from compressed sensing [19, 59, 30, 33, 13, 27, 58, 26].
- In practice, the underlying density w may be unknown so that one must resort to data-driven approaches in which moments of w must be approximated with data [48, 24].
- On the applications side, least-squares procedures for polynomial approximation have great utility in parametric uncertainty quantification, which seeks to understand solutions to parametric partial differential equations [15, 1, 40, 11].

There are also several possible directions for new research that one can explore:

- We have discussed randomized least-squares procedures, which allow bad behavior with low probability. However, optimal deterministic constructions would not suffer from the "in high probability" caveat. We have mentioned some deterministic strategies in the previous list, but there does not yet exist consensus on the "best" deterministic strategy for high-dimensional approximation. In addition, many existing approaches construct an approximation grid via numerical optimization, which can be very expensive.

- A full understanding of the asymptotics of induced distributions is elusive. A concrete open problem is to establish the conjecture (8.1), but a more general goal is a comprehensive understanding of the large-degree asymptotics of such sampling densities. Possible avenues of exploration include recent results from pluripotential theory [8].
- Adaptive construction of polynomial (or nonpolynomial) approximation spaces is an ongoing area of research. While existing approaches are effective, they still succumb to the curse of dimensionality, and some practical issues in adaptivity are being investigated.

REFERENCES

[1] I. Babuška, F. Nobile, and R. Tempone, *A stochastic collocation method for elliptic partial differential equations with random input data*, SIAM Rev., 52 (2010), pp. 317–355, https://doi.org/10.1137/100786356. (Cited on p. 505)

[2] J. Bäck, F. Nobile, L. Tamellini, and R. Tempone, *Stochastic spectral Galerkin and collocation methods for PDEs with random coefficients: A numerical comparison*, in Spectral and High Order Methods for Partial Differential Equations, Lect. Notes Comput. Sci. Eng. 76, Springer, Heidelberg, 2011, pp. 43–62. (Cited on p. 503)

[3] P. Binev, A. Cohen, W. Dahmen, and R. DeVore, *Universal algorithms for learning theory.* II. *Piecewise polynomial functions*, Constr. Approx., 26 (2007), pp. 127–152. (Cited on p. 484)

[4] P. Binev, A. Cohen, W. Dahmen, R. DeVore, and V. Temlyakov, *Universal algorithms for learning theory.* I. *Piecewise constant functions*, J. Mach. Learn. Res., 6 (2005), pp. 1297–1321. (Cited on p. 484)

[5] T. Bloom and N. Levenberg, *Weighted pluripotential theory in C^N*, Amer. J. Math., 125 (2003), pp. 57–103, https://doi.org/10.1353/ajm.2003.0002. (Cited on p. 501)

[6] L. Bos, J.-P. Calvi, N. Levenberg, A. Sommariva, and M. Vianello, *Geometric weakly admissible meshes, discrete least squares approximations and approximate Fekete points*, Math. Comp., 80 (2011), pp. 1623–1638. (Cited on p. 505)

[7] L. Bos, S. De Marchi, A. Sommariva, and M. Vianello, *Computing multivariate Fekete and Leja points by numerical linear algebra*, SIAM J. Numer. Anal., 48 (2010), pp. 1984–1999, https://doi.org/10.1137/090779024. (Cited on p. 505)

[8] L. Bos and N. Levenberg, *Bernstein-Walsh theory associated to convex bodies and applications to multivariate approximation theory*, Comput. Methods Function Theory, 18 (2018), pp. 361–388, https://doi.org/10.1007/s40315-017-0220-4. (Cited on p. 506)

[9] R. Brook and G. Arnold, *Applied Regression Analysis and Experimental Design*, Chapman & Hall, 2018. (Cited on p. 484)

[10] J.-P. Calvi and N. Levenberg, *Uniform approximation by discrete least squares polynomials*, J. Approx. Theory, 152 (2008), pp. 82–100, https://doi.org/10.1016/j.jat.2007.05.005. (Cited on p. 505)

[11] A. Chkifa, A. Cohen, G. Migliorati, F. Nobile, and R. Tempone, *Discrete least squares polynomial approximation with random evaluations—application to parametric and stochastic elliptic PDEs*, ESAIM Math. Model. Numer. Anal., 49 (2015), pp. 815–837. (Cited on p. 505)

[12] A. Chkifa, A. Cohen, and C. Schwab, *High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs*, Found. Comput. Math., 14 (2014), pp. 601–633, https://doi.org/10.1007/s10208-013-9154-z. (Cited on p. 505)

[13] A. Chkifa, N. Dexter, H. Tran, and C. Webster, *Polynomial approximation via compressed sensing of high-dimensional functions on lower sets*, Math. Comp., 87 (2018), pp. 1415–1450, https://doi.org/10.1090/mcom/3272. (Cited on p. 505)

[14] A. Cohen, M. A. Davenport, and D. Leviatan, *On the stability and accuracy of least squares approximations*, Found. Comput. Math., 13 (2013), pp. 819–834, https://doi.org/10.1007/s10208-013-9142-3. (Cited on pp. 485, 496, 505)

[15] A. Cohen, R. DeVore, and C. Schwab, *Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs*, Found. Comput. Math., 10 (2010), pp. 615–646, https://doi.org/10.1007/s10208-010-9072-2. (Cited on p. 505)

[16] A. Cohen and G. Migliorati, *Optimal weighted least-squares methods*, SMAI J. Comput. Math., 3 (2017), pp. 181–203, https://doi.org/10.5802/smai-jcm.24. (Cited on pp. 485, 498, 499, 500, 501, 505)

[17] A. Cohen, G. Migliorati, and F. Nobile, *Discrete least-squares approximations over optimized downward closed polynomial spaces in arbitrary dimension*, Constr. Approx., 45 (2017), pp. 497–519, https://doi.org/10.1007/s00365-017-9364-8. (Cited on p. 505)

[18] J. Dick and F. Pillichshammer, *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*, Cambridge University Press, New York, 2010. (Cited on p. 505)

[19] A. Doostan and H. Owhadi, *A non-adapted sparse approximation of PDEs with stochastic inputs*, J. Comput. Phys., 230 (2011), pp. 3015–3034, https://doi.org/10.1016/j.jcp.2011.01.002. (Cited on pp. 484, 505)

[20] N. R. Draper and H. Smith, *Applied Regression Analysis*, 3rd ed., Wiley, 1998. (Cited on p. 484)

[21] V. V. Fedorov, *Theory of Optimal Experiments*, Academic Press, 1972. (Cited on p. 484)

[22] W. Gautschi, *Orthogonal Polynomials: Computation and Approximation*, Oxford University Press, 2004. (Cited on p. 488)

[23] W. Gautschi and S. Li, *A set of orthogonal polynomials induced by a given orthogonal polynomial*, Aequationes Math., 46 (1993), pp. 174–198, https://doi.org/10.1007/BF01834006. (Cited on p. 499)

[24] L. Guo, Y. Liu, and T. Zhou, *Data-driven polynomial chaos expansions: A weighted least-square approximation*, J. Comput. Phys., 381 (2019), pp. 110–128. (Cited on pp. 501, 505)

[25] L. Guo, A. Narayan, L. Yan, and T. Zhou, *Weighted approximate Fekete points: Sampling for least-squares polynomial approximation*, SIAM J. Sci. Comput., 40 (2018), pp. A366–A387, https://doi.org/10.1137/17M1140960. (Cited on pp. 503, 505)

[26] L. Guo, A. Narayan, and T. Zhou, *A gradient enhanced ℓ_1-minimization for sparse approximation of polynomial chaos expansions*, J. Comput. Phys., 367 (2018), pp. 49–64. (Cited on p. 505)

[27] L. Guo, A. Narayan, T. Zhou, and Y. Chen, *Stochastic collocation methods via ℓ_1 minimization using randomized quadratures*, SIAM J. Sci. Comput., 39 (2017), pp. A333–A359, https://doi.org/10.1137/16M1059680. (Cited on p. 505)

[28] L. Gyørfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*, Springer-Verlag, Berlin, 2005. (Cited on p. 484)

[29] M. Hadigol and A. Doostan, *Least squares polynomial chaos expansion: A review of sampling strategies*, Comput. Methods Appl. Mech. Engrg., 332 (2018), pp. 382–407, https://doi.org/10.1016/j.cma.2017.12.019. (Cited on p. 505)

[30] J. Hampton and A. Doostan, *Compressive sampling of polynomial chaos expansions: Convergence analysis and sampling strategies*, J. Comput. Phys, 280 (2015), pp. 363–386. (Cited on p. 505)

[31] J. Hampton and A. Doostan, *Coherence motivated sampling and convergence analysis of least squares polynomial chaos regression*, J. Comput. Phys., 290 (2015), pp. 73–97. (Cited on p. 484)

[32] S. Hosder, R. Walters, and M. Balch, *Point-collocation nonintrusive polynomial chaos method for stochastic computational fluid dynamics*, AIAA J., 48 (2010), pp. 2721–2730. (Cited on p. 484)

[33] J. D. Jakeman, A. Narayan, and T. Zhou, *A generalized sampling and preconditioning scheme for sparse approximation of polynomial chaos expansions*, SIAM J. Sci. Comput, 39 (2017), pp. A1114–1144, https://doi.org/10.1137/16M1063885. (Cited on p. 505)

[34] J. Jobson, *Applied Multivariate Data Analysis: Regression and Experimental Design*, Springer, New York, 2012. (Cited on p. 484)

[35] M. Klimek, *Pluripotential Theory*, Oxford University Press, Oxford, 1991. (Cited on p. 501)

[36] M. Kutner, C. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, 5th ed., McGraw-Hill/Irwin, Boston, 2004. (Cited on p. 484)

[37] C. Lemieux, *Monte Carlo and Quasi-Monte Carlo Sampling*, Springer, New York, 2009, https://doi.org/10.1007/978-0-387-78165-5. (Cited on p. 505)

[38] G. Migliorati, *Adaptive polynomial approximation by means of random discrete least squares*, in Numerical Mathematics and Advanced Applications—ENUMATH 2013, A. Abdulle, S. Deparis, D. Kressner, F. Nobile, and M. Picasso, eds., Lect. Notes Comput. Sci. Eng. 103, Springer, Cham, 2015, pp. 547–554. (Cited on p. 505)

[39] G. Migliorati and F. Nobile, *Analysis of discrete least squares on multivariate polynomial spaces with evaluations at low-discrepancy point sets*, J. Complexity, 31 (2015), pp. 517–542. (Cited on pp. 503, 505)

[40] G. Migliorati, F. Nobile, E. von Schwerin, and R. Tempone, *Approximation of quantities of interest in stochastic PDEs by the random discrete L^2 projection on polynomial spaces*, SIAM J. Sci. Comput., 35 (2013), pp. A1440–A1460, https://doi.org/10.1137/120897109. (Cited on pp. 484, 505)

[41] A. Narayan, *Induced Distributions*, https://github.com/akilnarayan/induced-distributions, 2017. (Cited on pp. 485, 501)

[42] A. Narayan, *Computation of induced orthogonal polynomial distributions*, Electron. Trans. Numer. Anal., 50 (2018), pp. 71–97, https://doi.org/10.1553/etna_vol50s71. (Cited on pp. 485, 501)

[43] A. Narayan, *Induced Distribution Examples*, https://github.com/akilnarayan/induced-distributions-examples, 2018. (Cited on p. 485)

[44] A. Narayan and J. D. Jakeman, *Adaptive Leja sparse grid constructions for stochastic collocation and high-dimensional approximation*, SIAM J. Sci. Comput., 36 (2014), pp. A2952–A2983, https://doi.org/10.1137/140966368. (Cited on p. 500)

[45] A. Narayan, J. D. Jakeman, and T. Zhou, *A Christoffel function weighted least squares algorithm for collocation approximations*, Math. Comp., 86 (2017), pp. 1913–1947. (Cited on pp. 485, 500, 501)

[46] A. Narayan and D. Xiu, *Stochastic collocation methods on unstructured grids in high dimensions via interpolation*, SIAM J. Sci. Comput., 34 (2012), pp. A1729–A1752, https://doi.org/10.1137/110854059. (Cited on p. 505)

[47] A. Narayan and T. Zhou, *Stochastic collocation on unstructured multivariate meshes*, Commun. Comput. Phys., 18 (2015), pp. 1–36. (Cited on pp. 484, 505)

[48] S. Oladyshkin and W. Nowak, *Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion*, Reliab. Eng. Syst. Saf., 106 (2012), pp. 179–190. (Cited on pp. 501, 505)

[49] E. Saff and V. Totik, *Logarithmic Potentials with External Fields*, Springer, Berlin, 1997. (Cited on p. 501)

[50] Y. Shin, K. Wu, and D. Xiu, *Sequential function approximation with noisy data*, J. Comput. Phys., 371 (2018), pp. 363–381, https://doi.org/10.1016/j.jcp.2018.05.042. (Cited on p. 505)

[51] Y. Shin and D. Xiu, *Nonadaptive quasi-optimal points selection for least squares linear regression*, SIAM J. Sci. Comput., 38 (2016), pp. A385–A411, https://doi.org/10.1137/15M1015868. (Cited on p. 505)

[52] T. Strohmer and R. Vershynin, *A randomized Kaczmarz algorithm with exponential convergence*, J. Fourier Anal. Appl., 15 (2008), art. 262, https://doi.org/10.1007/s00041-008-9030-4. (Cited on p. 505)

[53] T. Tang and T. Zhou, *On discrete least-squares projection in unbounded domain with random evaluations and its application to parametric uncertainty quantification*, SIAM J. Sci. Comput., 36 (2014), pp. A2272–A2295, https://doi.org/10.1137/140961894. (Cited on p. 484)

[54] L. N. Trefethen, *Multivariate polynomial approximation in the hypercube*, Proc. Amer. Math. Soc., 145 (2017), pp. 4837–4844, https://doi.org/10.1090/proc/13623. (Cited on p. 490)

[55] L. N. Trefethen, *Approximation Theory and Approximation Practice*, SIAM, Philadelphia, 2012. (Cited on p. 484)

[56] J. A. Tropp, *User-friendly tail bounds for sums of random matrices*, Found. Comput. Math., 12 (2012), pp. 389–434, https://doi.org/10.1007/s10208-011-9099-z. (Cited on p. 497)

[57] D. Xiu and G. E. Karniadakis, *The Wiener–Askey polynomial chaos for stochastic differential equations*, SIAM J. Sci. Comput., 24 (2002), pp. 619–644, https://doi.org/10.1137/S1064827501387826. (Cited on p. 484)

[58] Z. Xu and T. Zhou, *A gradient enhanced ℓ_1 approach for the recovery of sparse trigonometric polynomials*, Commun. Comput. Phys., 24 (2018), pp. 286–308. (Cited on p. 505)

[59] L. Yan, L. Guo, and D. Xiu, *Stochastic collocation algorithms using ℓ_1-minimization*, Internat. J. Uncertainty Quantif., 2 (2012), pp. 279–293, https://doi.org/10.1615/Int.J.UncertaintyQuantification.2012003925. (Cited on p. 505)

[60] T. Zhou, A. Narayan, and D. Xiu, *Weighted discrete least-squares polynomial approximation using randomized quadratures*, J. Comput. Phys., 298 (2015), pp. 787–800. (Cited on p. 484)

[61] T. Zhou, A. Narayan, and Z. Xu, *Multivariate discrete least-squares approximations with a new type of collocation grid*, SIAM J. Sci. Comput., 36 (2014), pp. A2401–A2422, https://doi.org/10.1137/130950434. (Cited on pp. 503, 505)